(12) UK Patent Application (19) GB (11) 2 342 263 (13) A

(43) Date of A Publication 05.04.2000

(21) Application No 9828604.0

(22) Date of Filing 23.12.1998

(30) Priority Data
(31) 9841131 (32) 30.09.1998 (33) KR

(71) Applicant(s)
Daewoo Electronics Co., Ltd.
(Incorporated in the Republic of Korea)
541 5-Ga, Namdaemoon-Ro, Jung-Ku, Seoul,
Republic of Korea

(72) Inventor(s)
Tae-Beom Lim

(74) Agent and/or Address for Service
Page White & Farrer
54 Doughty Street, LONDON, WC1N 2LS,
United Kingdom

(51) INT CL[7]
H04N 7/173

(52) UK CL (Edition R)
H4R RCSS

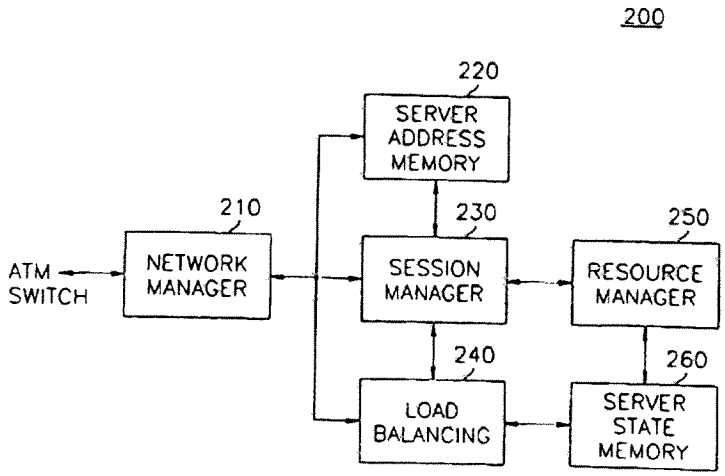
(56) Documents Cited
EP 0782337 A2 US 5737747 A US 5544327 A

(58) Field of Search
UK CL (Edition Q) H4K KOD3 , H4R RCSS RCST RCT RCX
INT CL[6] H04N 7/173
ONLINE - EPODOC, WPI

(54) Abstract Title
Server load balancing in a video on demand system

(57) In a video on demand system having N distributed servers (110 to 140, Fig. 1) and M users (410 to 440), a session and resource manager 200 balances the loading on the servers. The manager 200 has a server state memory 260 which stores addresses and bit stream pumping capability (maximum bandwidth of a bit stream which the server can provide) of each server, and a memory 220 stores the address of a default server which can be a server which was last accessed by a user and can be arbitrarily predetermined. In response to a VOD request from a user, a load balancing block 240 of manager 200 checks the address in memory 220 and then determines from memory 260 whether the default server has a maximum bandwidth which is greater than or equal to the bandwidth of a bit stream corresponding to the required VOD service. If the comparison result is affirmative, this server is assigned to the VOD request. If the comparison result is negative, the capability of a next server to provide the requested service is examined. When a server is found which is capable of providing the requested service, that server is assigned and the memory 220 is updated with the address of that server. If no server is found which is capable of providing the requested service, an error code is generated.

FIG. 2



GB 2 342 263 A

FIG. 1

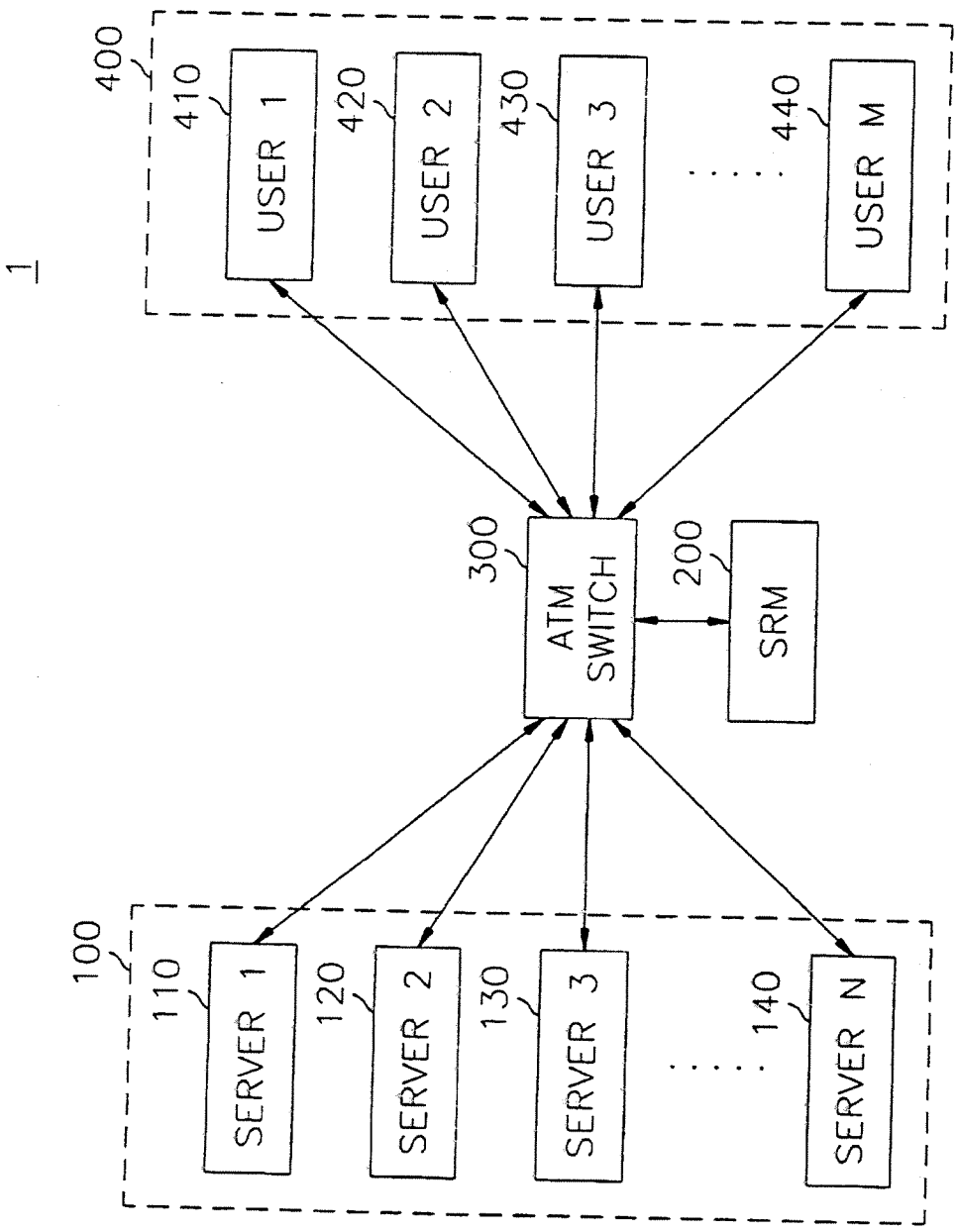

1
400
410
USER 1
420
USER 2
430
USER 3
440
USER M
300
ATM SWITCH
200
SRM
100
110
SERVER 1
120
SERVER 2
130
SERVER 3
140
SERVER N

FIG. 2

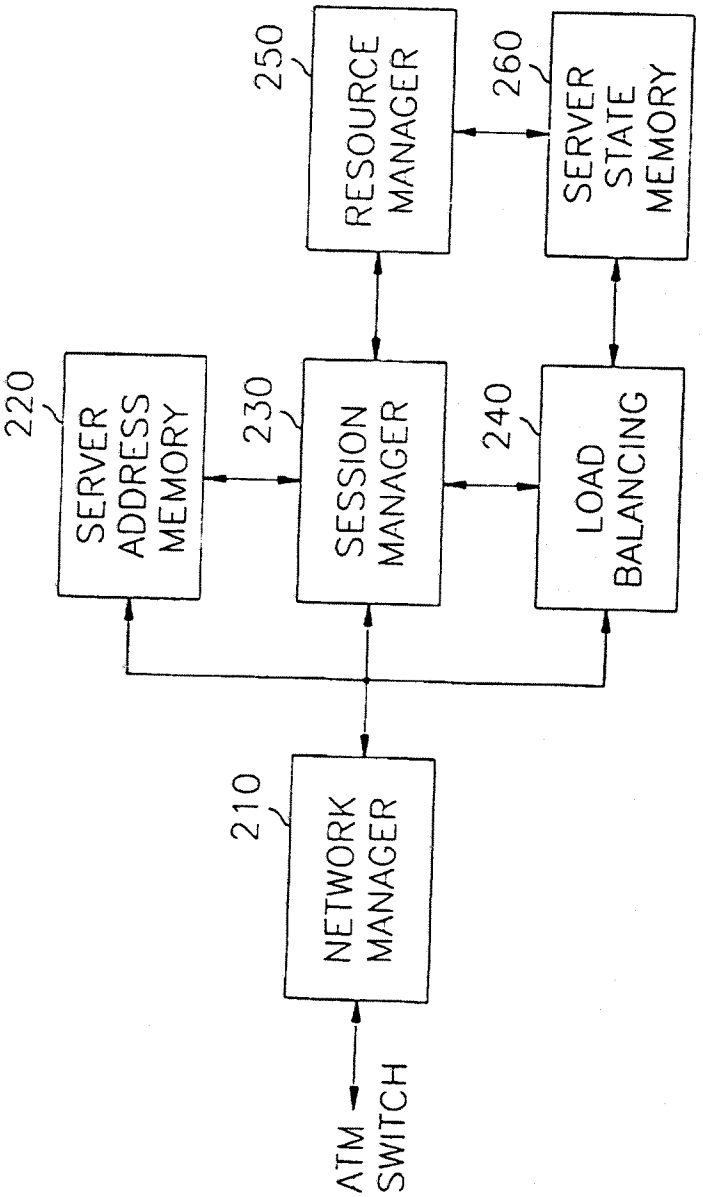
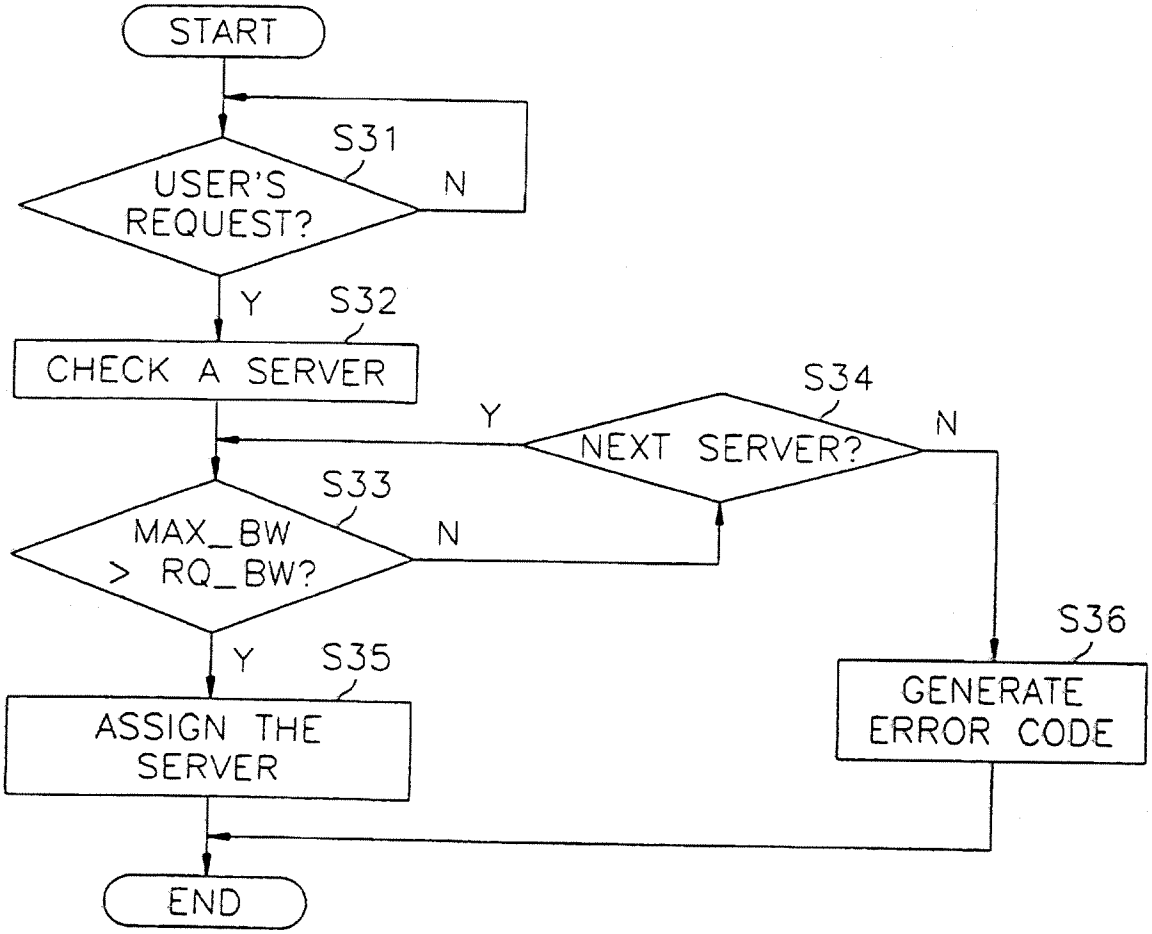

200
250
RESOURCE MANAGER
260
SERVER STATE MEMORY
220
SERVER ADDRESS MEMORY
230
SESSION MANAGER
240
LOAD BALANCING
210
NETWORK MANAGER
ATM SWITCH

FIG. 3


START
S31
USER'S REQUEST?
N
Y
S32
CHECK A SERVER
S34
NEXT SERVER?
Y
N
S33
MAX_BW > RQ_BW?
N
Y
S35
ASSIGN THE SERVER
S36
GENERATE ERROR CODE
END

LOAD BALANCING METHOD AND APPARATUS FOR USE IN A VIDEO ON DEMAND SYSTEM HAVING DISTRIBUTED SERVERS

The present invention relates to a method and apparatus for use in a video on demand system; and, more particularly, to a load balancing method and apparatus for use in a video on demand system having distributed servers.

As the so-called information superhighway is being developed, a wide bandwidth communication channel which interconnects households and businesses promises to provide many services to those connected to it. These services may include banking at home, instant access to large databases and real time interaction with virtual communities of people with common interests. Of the services available through the superhighway, one that has received a great deal of corporate and media attention is the supply of video on demand.

Desirable VOD services may include such videos as movies, sporting events, interactive games, home shopping, textual information, educational programs and arts programs. Videos generally include both video and audio portions, although a video may only have an image portion as in textual information, or only an audio portion as in, e.g., music.

In order to immediately provide users with requested

services, a VOD system comprises an information provider which is equipped with a server having a large capacity and a network which transmits information between the information provider and the users or between users. Although the server having the large capacity can provide a large amount of information at a time, it is very expensive and complicated to implement. Thus, there is a need for replacing a server having a large capacity with a multiplicity of servers having small capacities to provide a cheaper video on demand system capable of accommodating the user's various requirements.

It is, therefore, a primary object of the invention to provide a load balancing method and apparatus for use in a video on demand system having distributed servers, to accommodate various services.

In accordance with one aspect of the present invention, there is provided a load balancing method for use in a video on demand(VOD) system having N number of distributed servers, wherein a user applies a request for a VOD service, N being a positive integer, comprising the steps of: (a) setting a value of i to 1, i being an index for the a distributed server; (b) storing an address of an ith server; (c) finding the ith server corresponding to the stored address; (d) checking a pumping capability of the ith server; (e) assigning the ith server to provide the requested VOD service to the

user if the pumping capability thereof is sufficient to provide the requested VOD service; (f) increasing the value of i by 1, if otherwise and returning to the step (b) until the value of i becomes N; and (g) generating an error code if none of the N distributed servers have a sufficient pumping capability.

In accordance with another aspect of the present invention, there is provided a load balancing apparatus for use in a video on demand(VOD) system having N number of distributed servers, wherein a user applies a request for a VOD service, N being a positive integer, comprising: a block for storing an address of an ith server, i being an index for the distributed servers; a block for storing maximum bandwidths corresponding to N number of distributed servers; a block for checking a pumping capability of the ith server; and a block for assigning the ith server to provide the requested VOD service to the user if the pumping capability thereof is sufficient to provide the requested VOD service.

The above and other objects and features of the present invention will become apparent from the following description of preferred embodiments given in conjunction with the accompanying drawings, in which:

Fig. 1 represents a block diagram of a video on demand system having distributed servers;

Fig. 2 provides a detailed diagram of a SRM shown in Fig. 1; and

Fig. 3 illustrates a flow chart showing a load balancing process.

Referring to Fig. 1, there is illustrated a schematic block diagram of a video on demand(VOD) system 1 having distributed servers. The VOD system 1 comprises distributed servers 100, a SRM(session and resource manager) 200, an ATM (asynchronous transfer mode) switch 300 and a multiplicity of users 400. For the sake of simplicity, there are shown only 4 servers, i.e., a server 1 110 to a server N 140 and 4 users, i.e., a user 1 410 to a user M 440, in Fig. 1, wherein N and M is a positive integer, respectively.

A user applies a request for a VOD service by means of a user's device, wherein each of user's devices is a conventional user device with which the user communicates with the distributed servers 100 through an ATM network. When the request is applied to the ATM switch 300, a configuration request requiring network parameters, such as an address of the corresponding server and a required bandwidth of the requested VOD service to be used in a session setup, is provided to the SRM 200.

Then, a configuration confirmation providing the network parameters, in response to the configuration request, is

generated by the SRM 200 and provided to the user through the ATM switch 300. The ATM switch 300 sets up a session between the user and the corresponding server and the server provides the requested VOD service to the user. The detailed structure and operation of the SRM 200 will be illustrated with reference to Figs. 2 and 3.

Referring to Fig. 2, a detailed block diagram of the SRM 200 shown in Fig. 1 is depicted, wherein the SRM 200 includes a network manager 210, a server address memory 220, a session manager 230, a load balancing block 240, a resource manager 250 and a server state memory 260.

The network manager 210 transmits messages between users 400 and distributed servers 100. That is, the network manager 210 processes a request from a user and information from one of the distributed servers 100 corresponding thereto; and provides the address of the corresponding server to the user through the ATM switch 300.

The server address memory 220 is initialized with an address of a default server. The default server can be a server which is accessed last by a user and can be arbitrarily predetermined. The server address memory 220 provides the address of the default server to the session manager 230 if the default server is determined to be capable of providing the requested VOD service by the load balancing block 240; and updates its contents with an address of a corresponding server and provides the address of the corresponding server to the

session manager 230 if otherwise, wherein the corresponding server is determined by the load balancing block 240.

The session manager 230 is provided with the address of the default server or the address of the corresponding server from the server address memory 220 and provides same to the user through the network manager 210. The session manager 230 sets up the session between the user and the default server or the corresponding server based on the provided address to thereby transmit and receive data between the user and the server therethrough, and disconnects the session after the server completes providing the requested VOD service to the user.

The load balancing block 240 checks whether the default server or the corresponding server can provide the required VOD service or not. For this, the load balancing block 240 extracts data representing a bit stream pumping capability of the server, i.e., a maximum bandwidth of a bit stream which the server can provide, and compares it with a bandwidth of a bit stream corresponding to the required VOD service. If the maximum bandwidth of the bit stream which the server can provide is greater than or equal to the bandwidth of the bit stream corresponding to the requested VOD service, the load balancing block 240 provides a control signal of a first level to the server address memory 220 to maintain the address of the server. And if otherwise, the load balancing block 240 provides a control signal of a second level to the server

address memory 220 to update its contents with an address of another server which is capable of providing the requested VOD service.

The resource manager 250 assigns network resources needed for the session manager 230 to set up the session. And, the server state memory 260 stores an address and a bit stream pumping capability of each server.

With reference to Fig. 3, a load balancing process of the SRM 200 is described. At step S31, it will be checked if a user's request for a VOD service is applied through the ATM switch 300. If the checked result is affirmative, the procedure goes to step S32; and if otherwise, the checking procedure will be continued at step S31. At step S32, the address stored at the server address memory 220 is checked and goes to step S33, wherein the maximum bandwidth MAX_BW of the server corresponding to the address and a bandwidth RQ_BW of the required VOD service is compared with each other.

If MAX_BW is greater than or equal to RQ_BW, the procedure goes to step S35; and if otherwise, the procedure goes to step S34. At step S35, the server corresponding to the address is assigned to the user to thereby provide the requested VOD service. And, at step S34, it is examined if a next server exists. If the examined result is affirmative, the procedure goes to step S33; and if otherwise, the procedure goes to step S36, wherein an error code indicating that there is no server capable of providing the required VOD

service is generated.

In this way, a server capable of providing the requested VOD service is detected and the address of the server is stored at the server address memory 220. The session manager 230 sets up a session between the user and the server and a bit stream corresponding to the requested VOD service is provided from the server to the user through the session.

In accordance with the present invention, distributed servers replace a server having a large capacity and sessions are set up based on the load balancing result by the SRM 200, to thereby achieve an implementation of a highly efficient operation with a substantial cost reduction.

While the present invention has been described with respect to certain preferred embodiments only, other modifications and variations may be made without departing from the scope of the present invention as set forth in the following claims.

Claims:

1. A load balancing method for use in a video on demand(VOD) system having N number of distributed servers, wherein a user applies a request for a VOD service, N being a positive integer, comprising the steps of:

(a) setting a value of i to 1, i being an index for the a distributed server;

(b) storing an address of an ith server;

(c) finding the ith server corresponding to the stored address;

(d) checking a pumping capability of the ith server;

(e) assigning the ith server to provide the requested VOD service to the user if the pumping capability thereof is sufficient to provide the requested VOD service;

(f) increasing the value of i by 1, if otherwise and returning to the step (b) until the value of i becomes N; and

(g) generating an error code if none of the N distributed servers have a sufficient pumping capability.

2. The method of claim 1, wherein the step (d) determines that the ith server has the sufficient pumping capability if a maximum bandwidth of a VOD service which the ith server can provide is greater than or equal to the bandwidth of the requested VOD service.

3. The method of claim 2, wherein the step (e) includes the steps of:

(e1) transmitting the address of the ith server to the user;

(e2) making a session between the ith server and the user; and

(e3) providing data corresponding to the requested VOD service to the user through the session.

4. A load balancing apparatus for use in a video on demand(VOD) system having N number of distributed servers, wherein a user applies a request for a VOD service, N being a positive integer, comprising:

means for storing an address of an ith server, i being an index for the distributed servers;

means for storing maximum bandwidths corresponding to N number of distributed servers;

means for checking a pumping capability of the ith server; and

means for assigning the ith server to provide the requested VOD service to the user if the pumping capability thereof is sufficient to provide the requested VOD service.

5. The apparatus of claim 4, wherein the checking means includes:

means for extracting a maximum bandwidth corresponding

to the ith server from the maximum bandwidth storing means, wherein the maximum bandwidth corresponding to the ith server is a maximum bandwidth of a VOD service which the ith server can provide;

means for comparing the maximum bandwidth corresponding to the ith server with a bandwidth of the requested VOD service; and

means for determining that the ith server has a sufficient pumping capability if the maximum bandwidth corresponding to the ith server is greater than or equal to the bandwidth of the requested VOD service.

6. The apparatus of claim 5, wherein the assigning means includes:

means for transmitting the address of the ith server to the user;

means for making a session between the ith server and the user; and

means for providing data corresponding to the requested VOD service to the user through the session.

7. The apparatus of claim 6, wherein the address storing means updates its contents with an address of a server which has the sufficient pumping capability.

8. A method, substantially as herein described with reference

to or as shown in figures 1 to 3 of the accompanying drawings.

9. An apparatus, constructed and arranged substantially as herein described with reference to or as shown in figures 1 to 3 of the accompanying drawings.

Application No: GB 9828604.0
Claims searched: 1 to 9

Examiner: M J Billing
Date of search: 25 May 1999

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.Q): H4K KOD3; H4R RCSS, RCST, RCT, RCX.

Int Cl (Ed.6): H04N 7/173.

Other: ONLINE - EPODOC, WPI.

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
A	EP0782337A2	(A T & T) - Fig.5, column 8 lines 33-46	1,4
X	US5737747	(EMC) - Figs.17-20; column 23 line 20 to column 25 line 12	1,4 at least
A	US5544327	(IBM) - Fig.8; column 7 lines 29-35	1,4

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.